BMC Evolutionary Biology

BioMed Central

Research article

Open Access

Genome-wide comparative phylogenetic analysis of the rice and Arabidopsis Dof gene families

Diego Lijavetzky*, Pilar Carbonero and Jesús Vicente-Carbajosa

Address: Laboratorio de Bioquímica y Biología Molecular, Departamento de Biotecnología-UPM, E.T.S. Ingenieros Agrónomos, Ciudad Universitaria s/n, Madrid 28040 SPAIN

Email: Diego Lijavetzky* - dlijavetzky@bit.etsia.upm.es; Pilar Carbonero - pcarbonero@bit.etsia.upm.es; Jesús Vicente-Carbajosa - jvicente@bit.etsia.upm.es

* Corresponding author

Published: 23 July 2003

BMC Evolutionary Biology 2003, **3**:17

Received: 22 April 2003

Accepted: 23 July 2003

This article is available from: http://www.biomedcentral.com/1471-2148/3/17

Abstract

Background: Dof proteins are a family of plant-specific transcription factors that contain a particular class of zinc-finger DNA-binding domain. Members of this family have been found to play diverse roles in gene regulation of processes restricted to the plants. The completed genome sequences of rice and Arabidopsis constitute a valuable resource for comparative genomic analyses, since they are representatives of the two major evolutionary lineages within the angiosperms. In this framework, the identification of phylogenetic relationships among Dof proteins in these species is a fundamental step to unravel functionality of new and yet uncharacterised genes belonging to this group.

Results: We identified 30 different Dof genes in the rice *Oryza sativa* genome and performed a phylogenetic analysis of a complete collection of the 36-reported *Arabidopsis thaliana* and the rice Dof transcription factors identified herein. This analysis led to a classification into four major clusters of orthologous genes and showed gene loss and duplication events in Arabidopsis and rice, that occurred before and after the last common ancestor of the two species.

Conclusions: According to our analysis, the Dof gene family in angiosperms is organized in four major clusters of orthologous genes or subfamilies. The proposed clusters of orthology and their further analysis suggest the existence of monocot specific genes and invite to explore their functionality in relation to the distinct physiological characteristics of these evolutionary groups.

Background

Detailed analyses of completely sequenced genomes reveal that a significant percentage of the encoded proteins corresponds to transcription factors (TF). These can be classified into several gene families according to the presence of particular DNA binding domains [1–8]. However, the analysis of a particular transcription factor should be done in the context of the family, to which it belongs, taking into account that functional redundancy is a very frequent event within eukaryotic TFs [9]. Moreover, transcription factors operate in complex networks based on protein-protein interactions and are often organized into regulatory cascades. Due to their crucial role in the regulation of gene expression, the study of TFs is of outstanding interest, and for the reasons stated above it should be ideally done from a genomic perspective [10,11].

The complete genomic sequence of *Arabidopsis thaliana* [3] and the shotgun-quality genomic sequence of *Oryza sativa*

[12–15] have been recently obtained, each constituting a model plant for a dicotyledonous and monocotyledonous species respectively. In both the Arabidopsis and rice genomes several groups of plant-specific TF have been described that are of great interest, since they may be involved in the regulation of events restricted to the plant kingdom [11]. One of these groups is the Dof (DNA binding with one finger) family, a particular class of zinc finger domain TFs [16,17] characterized by a conserved region of 50 amino acids with a C_2-C_2 finger structure, associated to a basic region, that binds specifically to DNA sequences with a 5'-T/AAAAG-3' core [18]. Dof proteins have been reported to participate in the regulation of gene expression in processes such as seed storage protein synthesis in developing endosperm [19,20], light regulation of genes involved in carbohydrate metabolism [21], plant defense mechanisms [22], seed germination [23–25], gibberellin response in post-germinating aleurone [26,27], auxin response [28–30] and stomata guard cell specific gene regulation [31].

Because hierarchy organization of genes reflects an ancient process of gene duplication and divergence, many of the theoretical and analytical tools of the phylogenetic systematics can be utilized in comparative genomics [5]. Here, this analytical approach, successfully applied in Arabidopsis [32], was used to perform a phylogenetic characterization of all the Arabidopsis and rice Dof transcription factors. As a first step we have revised and annotated all the rice Dof genes and compared them with those from Arabidopsis. This phylogenetic analysis, led to the definition of four clusters of rice and Arabidopsis orthologous genes and to identify the minimum complement of Dof genes that were present in the angiosperm common ancestor. We also discuss on the relevance to recognize groups of orthology as the basis for further characterization of Dof genes of unknown function.

Results

Identification of a comprehensive set of Dof proteins from rice and Arabidopsis

Rice Dof genes

Oryza sativa ssp. *japonica* and *O. sativa* ssp. *indica* draft genome sequences were simultaneously released [13,14] and more recently the high quality finished sequences of *O. sativa* ssp. *japonica* chromosomes 1 and 4 [33,34]. We analyzed both genomes in order to assemble a complete and non-redundant set of rice Dof genes. The nucleotide and deduced amino acid sequences of the Dof domains were used to perform independent Blast searches [35] through several rice databases: Rice TIGR db, DDBJ and TMRI Rice Genome Database (for the *japonica* genome) and the NCBI *O. sativa* BLAST page (for the *indica* genome). A total of 30 non-redundant Dof transcription factors were identified in *japonica* and *indica*. Among them, twenty-seven sequences were almost identical in both species, while two *japonica* genes were not clearly identified in *indica* and one *indica* gene was not clearly identified in *japonica*. To explore this discrepancy we used the complete sequences of the *japonica* genes, partially detected in *indica*, to perform a BLAST search in the *indica* database and *vice versa*. Portions of the two genes partially detected in *indica* (OsDof-5 and OsDof-20) were found in the corresponding database. In the case of OsDof-5, a DNA fragment 3' to the Dof domain was identified as a perfect match to the *japonica* gene used as the query sequence. This fragment corresponds to a terminal region of a short contig assembly (AAAA01006178.1). The other *indica* sequence, OsDof-20, was found within a misassembled contig (AAAA01062012.1) as a completely homologous match to the *japonica* sequence. For the *indica* gene not clearly identified in *japonica* (OsDof-30), a DNA fragment 5' to the Dof domain was identified as a perfect match against the *indica* gene used as the query sequence. This fragment corresponds to a terminal region of a 6 kb-contig assembly at the TMRI rice genome project (CL037947.70). Since none of the *indica* and TMRI *japonica* contigs have been mapped to a rice genome, we were unable to provide the chromosome location for OsDof-29 and OsDof-30 in Table 1. Gene structure and the corresponding deduced amino acid sequences for all the *indica* Dof TFs and eight from *japonica* were processed with the help of the RiceGAAS annotation system. The remaining *japonica* genes were obtained from the rice TIGR annotation database. According to the predicted structures, approximately half of the rice Dof TFs (16) contains one or more introns (Table 1).

Arabidopsis Dof genes

A non-redundant and complete compilation of the Arabidopsis Dof genes was obtained from the *At* TIGR db and MIPS MATDB databases. A total of 36 annotated TFs belonging to the Dof gene family were extracted from these sources (Table 2). In a previous publication, Riechmann [11] indicated the existence of 37 Dof encoding genes in Arabidopsis. However, the presence of several stop codons within the ORF of At1g65935 suggests that it is most probably a pseudogene [36] and was therefore excluded from our analysis. Structural examination of the remaining 36 genes revealed the presence of introns in half of the sequences, generally placed upstream of the Dof domain. Of those, 15 contained just one intron (Table 2).

Phylogenetic analysis and recognition of Dof families in rice and Arabidopsis

In order to evaluate the evolutionary relationship among the rice Dof proteins, we performed a phylogenetic analysis based on their DNA binding domain sequences (Figure 1). Pair-wise amino acid similarities were higher than

Table 1: Rice Dof transcription factors

Sequence Name[a]	*Japonica* BAC/PAC Name[b]	*Indica* Contig Name[c]	Predicted Gene structure[d]		Chomosome[e]	Group[f]
OsDof-1	P0505D12	AAAA01001390.1	___Dof___	(T)	1	c_1
OsDof-2	P0453A06	AAAA01021216.1	_▼_Dof___	(T)	1	d_1
OsDof-3	P0001B06	AAAA01003833.1	___Dof___	(T)	1	d_2
OsDof-4	P0038F12	AAAA01000894.1	_$▼_2$_Dof___	(T)	1	d_1
OsDof-5	P0007F06	AAAA01006178.1[i]	___Dof___	(T)	1	d_1
OsDof-6	B1131G08	AAAA01001653.1	___Dof___	(R)	1	d_1
OsDof-7	49D11	AAAA01000209.1	_▼_Dof___	(T)	2	c_1
OsDof-8	B1121A12	AAAA01004110.1	_$▼_2$_Dof_▼_	(T)	2	a
OsDof-9	P0657H12	AAAA01000342.1	___Dof___	(T)	2	c_2
OsDof-10	OSJNBa0009N02	AAAA01004826.1	___Dof_$▼_2$_	(T)	2	c_2
OsDof-11	OSJNBa0010D22	AAAA01014987.1	___Dof___	(T)	3	a
OsDof-12	OSJNBa0091P11	AAAA01007635.1	_▼_Dof___	(T)	3	d_1
OsDof-13	OSJNBa0063J18	AAAA01000005.1	___Dof_▼_	(T)	3	d_3
OsDof-14	OSJNBa0002D01	AAAA01000212.1	_▼_Dof___	(T)	3	b
OsDof-15	OSJNBa0079B15	AAAA01003692.1	___Dof___	(T)	3	a
OsDof-16	OJ1754_E06	AAAA01000954.1	___Dof___	(R)	3	b
OsDof-17	OSJNBa0064G10	AAAA01000011.1	_▼_Dof___	(T)	4	d_2
OsDof-18	OSJNB0005N02	AAAA01002346.1	___Dof_$▼_2$_	(T)	4	c_2
OsDof-19	P0016H04	AAAA01007236.1	_▼_Dof___	(T)	5	b
OsDof-20	P0491D10	AAAA01062012.1[j]	___Dof_▼_	(T)	6	a
OsDof-21	P0407H12	AAAA01008294.1	___Dof___	(T)	7	b
OsDof-22	P0483G08	AAAA01004849.1	___Dof___	(T)	7	b
OsDof-23	OSJNBa0060O17	AAAA01004298.1	___Dof___	(R)	7	d_1
OsDof-24	P0605H02	AAAA01033718.1	___Dof___	(R)	8	d_3
OsDof-25	P0556A05	AAAA01001409.1	_$▼_2$_Dof___	(R)	9	d_3
OsDof-26	OSJNBa0060A14	AAAA01002383.1	_$▼_3$_Dof___	(R)	10	d_1
OsDof-27	OSJNBa0066I08	AAAA01003749.1	_$▼_2$_Dof_$▼_2$_	(T)	10	d_2
OsDof-28	OSJNBa0016C14	AAAA01003609.1	_▼_Dof___	(T)	12	b
OsDof-29	CL012178.39[g]	AAAA01002638.1	___Dof___	(R)		c_1
OsDof-30	CL037947.70[h]	AAAA01068763.1	___Dof___	(R)		d_3

[a] Sequence name designation is arbitrary. [b] *O. sativa* ssp. *japonica* BAC/PAC clone name at the rice TIGR db. [c] *O. sativa* ssp. *indica* contig name at the NCBI *O. sativa* BLAST page. [d] Gene structure prediction according to the rice TIGR db (T) or to the RiceGAAS system (R). Intron (▼) relative position respect to the Dof domain. Sub-index indicates number of introns. [e] Chromosome assignment according to *O. sativa* ssp. *japonica* at the rice TIGR db. [f] Group designation after the phylogenetic analysis displayed in Figure 2B. [g] *O. sativa* ssp. *japonica* contig name at the rice TMRI db. [h] *O. sativa* ssp. *japonica* contig (at TMRI) containing incomplete gene (lacking the Dof domain) [i] *O. sativa* ssp. *indica* contig containing incomplete gene (lacking the Dof domain) [i] *O. sativa* ssp. *indica* contig containing a misassembled gene.

50%, a threshold conventionally used to classify a group of genes as a gene family [5,37]. Consistent with the un-rooted tree obtained by the neighbor-joining algorithm (Figure 2B) four groups were defined (a, b, c and d), two of which were further divided into subgroups supported by the presence and position of introns (Table 1), bootstrapping values and the occurrence of common protein motifs outside of the Dof domain (Figure 3 and Table 3).

An equivalent phylogenetic analysis of Dof domain sequences was done in Arabidopsis. The un-rooted tree inferred from the neighbor-joining analysis in displayed in Figure 2C. Our results show that the Arabidopsis Dof gene family can be organized into four groups or subfamilies (A, B, C and D). Groups B, C and D were further subdivided into subgroups, according the same criteria applied in the analysis of the rice Dof proteins. In order to detect putative duplicated genes in the Arabidopsis genome, we examined sequence redundancy between pairs of closely related Dof proteins. Using the Arabidopsis Redundancy Viewer (MATDB), we found ten pairs of genes (Table 1) on genomic regions associated with major genomic duplication events that occurred in Arabidopsis [38–40].

Comparison of the Arabidopsis and rice Dof proteins and determination of orthology relationships

To evaluate the evolutionary relationships within the Dof gene family, we performed a combined phylogenetic analysis of the 66 Arabidopsis and rice sequences to obtain a joint tree (Figure 2A). The tree topology, as well as the group and subgroup organization, resembled those from the rice and Arabidopsis individual trees (Figures 2B and 2C). The tree presented in Figure 2A identified putative

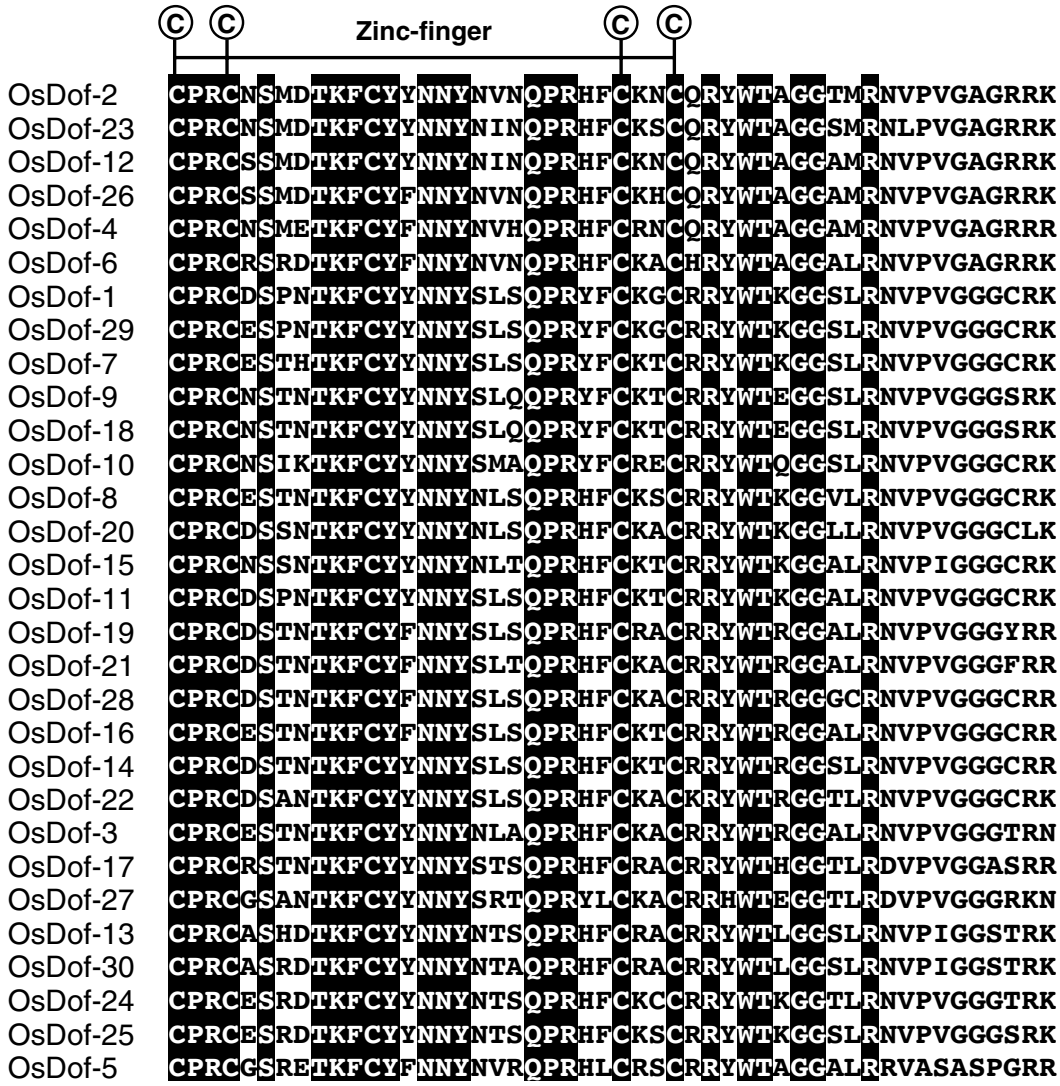

Figure 1
Dof domain sequence alignment of the annotated rice proteins. The four cysteine residues putatively responsible of the zinc-finger structure are indicated. Identical amino acids are highlighted in black. Gene names correspond to those listed in Table 1.

orthologs (i.e. OsDof-22/At1g28310 or OsDof-7/At5g62940), paralogs (i.e. OsDof-9/OsDof-18 or At2g46590/At3g61850), as well as presumed gene loss events (i.e. clusters C_3 or d_3). Four major clusters of orthologous groups (MCOG) were identified (*Aa*, *Bb*, *Cc* and *Dd*). A cluster of orthologous groups (COG) is defined as individual orthologous genes or orthologous groups of paralogous genes (i.e. cluster $C_{2.2}$+c_1) [1,41,42]. Most Arabidopsis and rice paralogous genes observed in Figure 2A were already displayed as paralogs in the respective trees (Figures 2B and 2C). Additionally, nearly all the Arabidopsis paralogs correspond to regions described as

Table 2: Arabidopsis Dof transcription factors

AGI code[a]	Predicted gene structure[b]	Chromosome	Group[c]	Gene Name[d]
At5g60850	___Dof___	5	A	*OBP4*
At1g51700[1]	___Dof___	1	A	
At3g21270[1]	___Dof___	3	A	
At1g07640	_▼_Dof___	1	B_1	*OBP2*
At3g55370[2]	___Dof_▼_	3	B_1	*OBP3*
At2g37590[2]	_▼_Dof___	2	B_1	
At2g28810	_▼_Dof_▼_	2	B_1	
At5g02460	_▼_Dof___	5	B_1	
At5g65590	___Dof___	5	B_2	
At4g38000	___Dof___	4	B_2	
At1g28310	___Dof___	1	B_2	
At2g28510	_▼_Dof___	2	C_1	
At3g45610[3]	_▼_Dof___	3	C_1	
At5g60200[3]	_▼_Dof___	5	C_1	
At5g62940	_▼_Dof___	5	C_1	
At1g64620	_▼_Dof___	1	$C_{2.1}$	
At4g00940	___Dof___	4	$C_{2.1}$	
At4g24060	_▼_Dof___	4	$C_{2.1}$	
At2g46590[4]	_▼_Dof___	2	$C_{2.1}$	*DAG2*
At3g61850[4]	_▼_Dof_▼_	3	$C_{2.1}$	*DAG1*
At3g52440	___Dof___	3	$C_{2.2}$	
At1g21340	___Dof___	1	$C_{2.2}$	
At4g21030[5]	___Dof___	4	C_3	
At4g21050[5]	___Dof___	4	C_3	
At4g21040[6]	___Dof___	4	C_3	
At4g21080[6]	___Dof___	4	C_3	
At2g34140[7]	___Dof___	2	D_1	
At1g29160[7]	___Dof___	1	D_1	*COG1*
At3g47500[8]	_▼_Dof___	3	D_1	*HPPBF*-2a
At5g62430[8]	_▼_Dof___	5	D_1	
At5g39660	_▼_Dof___	5	D_1	
At1g26790[9]	_$▼_2$_Dof___	1	D_1	
At1g69570[9]	_▼_Dof___	1	D_1	*HPPBF*-2b
At1g47650	___Dof___	1	D_2	
At5g66940[10]	___Dof___	5	D_2	
At3g50410[10]	___Dof___	3	D_2	*OBF*

[a] Names from the Arabidopsis Genome Initiative (AGI). Genes sharing the same superscript number were found in duplicated genomic regions according to the Redundancy Viewer (MIPS/MATDB). [b] Data from the *At* TIGR db. Intron (▼) relative position respect to the Dof domain. Sub-index indicates number of introns. [c] Group designation after the phylogenetic analysis displayed in Figure 2C. [d] Gene names originally utilized for publication or database registration.

genomic redundancies (Table 2). Just three genes were found in unexpected locations in the combined tree. A possible reason for this fact could be the presence of an apparent orthologous from the other species, generating a better support for the new location (i.e. OsDof-3 and At1g07640). On the other hand, relocation of At5g65590 seems to be an artifact of sequence similarities across the Dof domain, since according to additional information of the conserved motifs outside of the Dof domain, location within MCOG *Bb* is more consistent (Figures 2A and 3).

Comparative analyses of the complete amino acid sequences of the Dof proteins by BLOCKS and MEME software (Figure 3) are in agreement with those of the presented phylogenetic analysis, since several family and subfamily specific conserved motifs (Table 3) could be determined for each of the defined groups in Figure 2A. Moreover, an additional tree derived from the MEME results, analyzing the Dof domain plus all the conserved motifs described in Table 3, presented the same group and subgroup organization as the tree displayed in Figure 2A (data not shown).

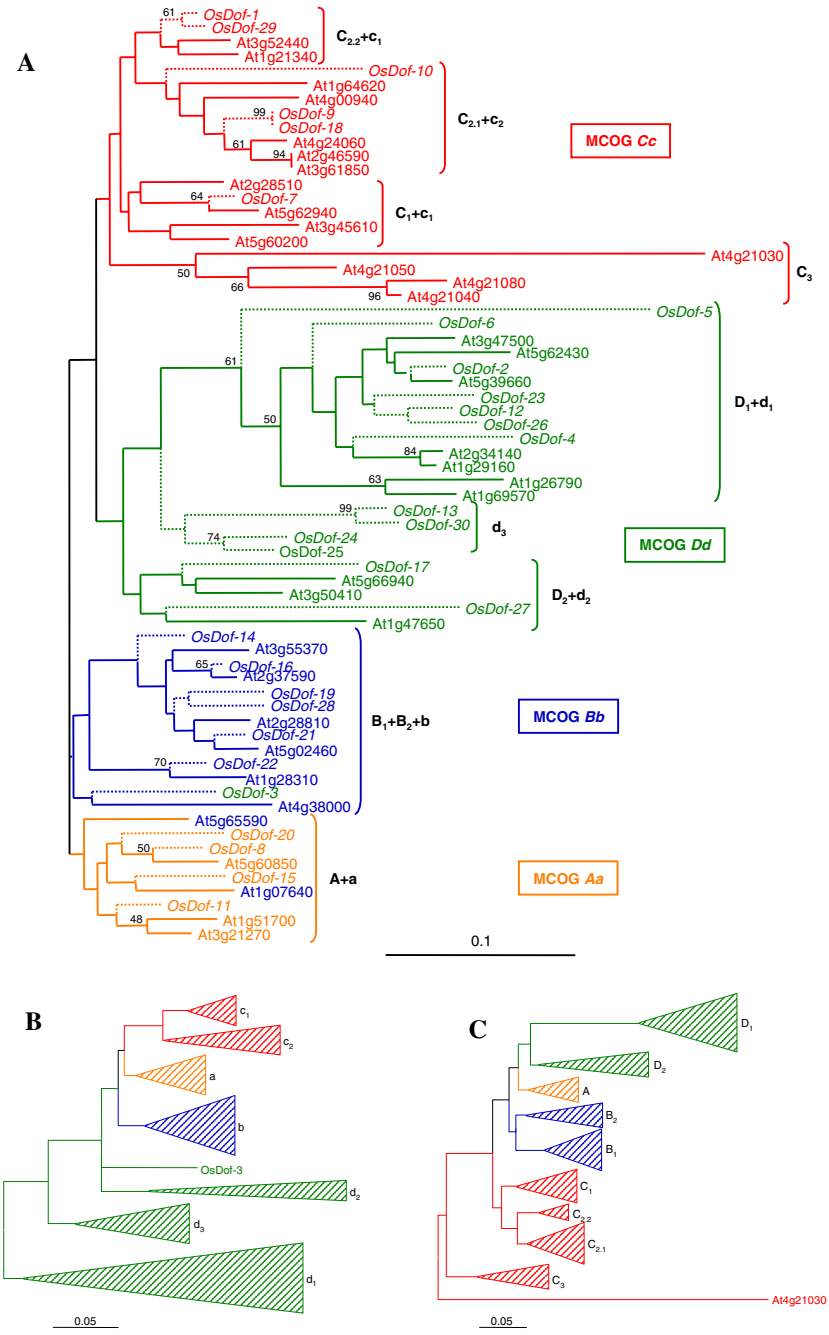

Figure 2

(A) Joined phylogenetic tree of the rice and Arabidopsis Dof gene families. The unrooted tree was inferred by the neighbor-joining method after the alignment of the Dof domain amino acid sequences of the 66 Arabidopsis and rice genes listed in Table 1 and 2 respectively. Arabidopsis (normal case) and rice genes (*italics*) are indicated at the end of solid and dotted branches, respectively. Bootstrapping values are indicated as percentages (when >50 %) along the branches. The resulting major clusters of orthologous genes (MCOG) are shown in different colors: *Aa* = orange, *Bb* = blue, *Cc* = red and *Dd* = green. Subscript numbers indicate the defined subgroups. The scale bar corresponds to 0.1 estimated amino acid substitution per site. **Phylogenetic trees of rice (B) and Arabidopsis (C) Dof gene families.** The unrooted trees were inferred by the neighbor-joining method after the alignment of the Dof domain amino acid sequences of the 30 rice and 36 Arabidopsis genes listed in Tables 1 and 2, respectively. The resulting groups are shown in different colors: A or a = orange, B or b = blue, C or c = red and D or d = green. Subscript numbers indicate the defined subgroups. The scale bar corresponds to 0.05 estimated amino acid substitutions per site. Genes belonging to the different groups are listed in Table 1 and 2.

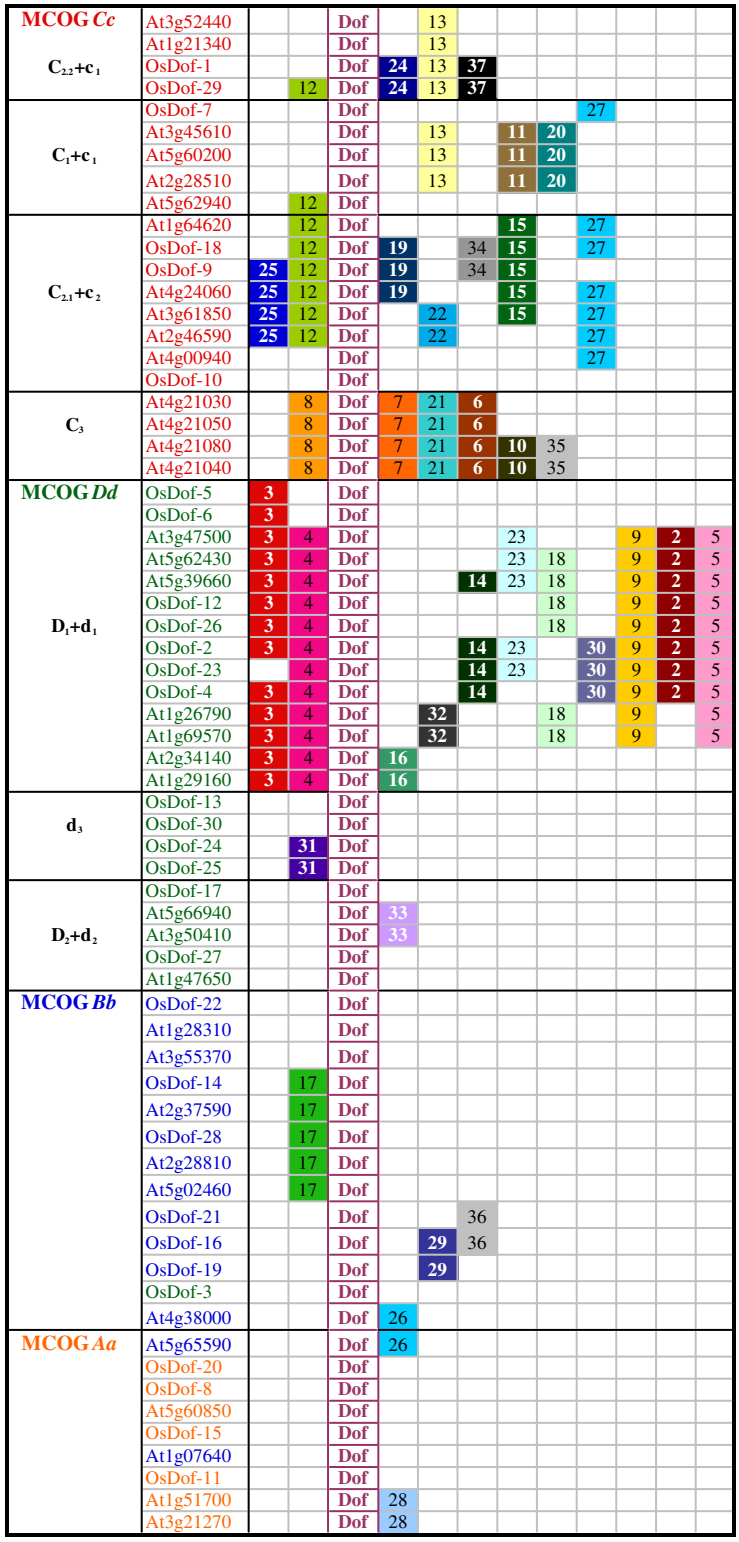

Figure 3

Schematic distribution of conserved motifs among the defined gene clusters in Figure 2A. Motifs were identified by means of the MEME software using the complete amino acid sequences of the 66 rice and Arabidopsis Dof genes listed in Tables 1 and 2. Positions of the identified motifs are relative to the Dof domain. Multilevel consensus sequences for the MEME defined motifs are listed in Table 3.

Table 3: Group and sub-group specific conserved motifs

Motif[a]	Multilevel consensus sequence[b]
1	*CPRCDSTNTKFCYYNNYSLSQPRHFCKACRRYWTKGGALRNVPVGGGCRK*[c]
2	KGEGCLWVPKTLRIDDPDEAAKSSIWTTLGIK
3	DDPGIKLFGKTIPF
4	KALKKPDKILP
5	LQANPAALSRSQNFQE
6	PMDRLAFGDESFEQDYYDVGSDDLIVNPLIGGS
7	KRAKIDQPSVAQMVSVEIQPGNHQPFFNVQENNDFVGSF
8	MDNLNVFANEDNQVNDVKPPP
9	SPTLGKHSRDE
10	YHMNPVDQFKWNQSFNNAMNMNYNN
11	RVLWGFPWQM
12	ERKARPQKDQ
13	IDLALAYAKFLKHH
14	ATALKFASDSPLCESMASVLDIGEK
15	RLLFPFEDLKPLVS
16	HGGFRHDFPMKRLRCYSDGQSC[d]
17	MVFSSVPAYLD
18	FYPVPAYWGC
19	KNPKLLHEGAQDLNLAFPHH
20	GHVDQIDSGREIW
21	VAAVGNHFGSLSEIHG
22	MMDSNSVLYSSLGFPTMPDYK
23	CFPGVPWPYTW
24	FGHRFHGPVRPDMILEGM
25	INVKPMEEI
26	IESLSCFNQDLHQKLQQQR
27	YWSGMI
28	WTDLAMNRAEK
29	LEQWRLPQIQQFPFFHAMDAM
30	WGCFSGWPNGAWNAPWI
31	EAGRRPAPQFAGVDLRRPKGY
32	NKGWPSSDHYLHITSEDCDN
33	HAAPIPATWQFEGLE
34	MELLRSTGCYM
35	MHPCHLEK
36	PIEFLGLPGNLQFW
37	DEEAKYDPFDSFPDDALSLHDCI

[a] Numbers correspond to the motifs described in Figure 3. [b] Sequences obtained from the analysis of the 66 rice and Arabidopsis Dof complete proteins with the MEME system. [c] Dof consensus sequence (in italics). [d] Predicted nuclear localization signal, according to Park et al [50] (underlined).

Discussion

Comparative genomic analysis of the rice and Arabidopsis Dof gene families

The main objective of this phylogenetic study was to identify putative orthologous and paralogous Dof genes, orthologs being defined as genes in different genomes that have been created by the splitting of taxonomic lineages, and paralogs as genes in the same genome created by gene duplication events [5,43]. Paralogs usually display different functions, while orthologs may retain the same function [1]. Distinguishing orthologous from paralogous genes is essential to comparative genomics. Indeed, the fundamental activity of comparative genomics is to track the presence, structural characteristics, function, and map position of orthologs in multiple genomes [5].

Considering the extensive annotation work done in Arabidopsis since the release of its sequenced genome [3], together with the analysis of rice sequences carried out in this study, we assume that most (or possibly all) of the Dof transcription factors from these species are represented in the 66 sequences documented (36 from Arabidopsis and 30 from rice). Our analyses of these sequences defined four MCOGs in rice and Arabidopsis (Figure 2A). Within each MCOG, particular clusters of paralogous and orthologous genes were identified, showing ancestral

duplication and gene loss events. These results were also corroborated through the construction of a rice/Arabidopsis reconciled tree [5,44] (data not shown). The tree presented (Figure 2A) showed considerable bootstrapping support for many of the defined groups and subgroups, but several clusters remained with poor supporting values. This fact was an expectable consequence of performing a study like the present with a 50 amino acid-length sequence, a constraint imposed by the lack of sequence conservation among Dof proteins outside this domain. However, it is worth to mention that most of the groups and subgroups defined were supported by additional criteria, such as gene structure and the presence of common protein motifs outside the Dof domain detected in the MEME analysis (Figure 3 & Table 3).

The Dof family in angiosperms: Rice (monocot) and Arabidopsis (dicot) specific genes

Although Dof proteins are exclusively found in the plant kingdom, searching EST databases allows to track the occurrence of Dof-encoding sequences from the unicellular algae *Clamydomonas*, to mosses and gymnosperms, indicating an ancient origin and the possibility of diversification throughout plant evolution. In this respect, comparisons of Dof repertoires from different organisms may give important insights into the evolutionary history of the family.

Assuming our compilation to be a complete catalog of the rice and Arabidopsis Dof transcription factors, we might postulate the existence of rice and Arabidopsis specific Dof genes, and by extension, putative monocot- and dicot-specific Dof genes. The genes belonging to Arabidopsis cluster C_3 and rice cluster d_3 (Figure 2A) might represent such situation, since each group has no apparent counterpart in the other species. To characterize these events further, we performed global searches (against whole plant nucleotide and protein sequences) with all the members of the two subfamilies (C_3, d_3). For this purpose, query sequences used in BLAST searches were selected outside the Dof domain, since this structure is highly conserved across the whole plant kingdom [36]. The sequence producing the lowest e value with the C_3 queries corresponds to the *Pisum sativum ERDP* gene (Accession BAA85655.1) followed by maize *PBF* and its orthologs from barley and wheat. *PBF*-like genes are Dof proteins known to participate in important regulatory processes of gene expression in the seeds of monocots [19,20]. This suggests an orthologous relationship between the Arabidopsis (dicot) subfamily C_3 and monocot *PBF*-like genes. This hypothesis is in agreement with our own experimental results, where At4g21080 shows seed specific expression (unpublished results).

BLAST searches with the putative monocot-specific d3 group identified two maize genes as the only sequences that were closely related. The maize genes *Dof1* and *Dof2* [21] are likely to be orthologous to OsDof-22/OsDof-16. These findings suggest the possible existence of monocot-specific genes (i.e. d_3 related), while no obvious dicot-specific genes were found based on the Arabidopsis sequences. Further analysis of Dof evolution will require the completion of genome projects currently underway and the isolation of Dof sequences, not available at present, from a broader spectrum of plant species.

Duplication events, gene function and phylogenic relationship

When comparing multi-gene families between species it is a common event to find several genes in one species that are collectively orthologs of a single gene in the other, indicating recent duplications exclusive to the former. In this situation, knowledge of gene function of certain members allows the confirmation of paralogous and orthologous relationships, otherwise difficult to infer merely from tree topologies. This is the case of *DAG1* (At3g61850) and *DAG2* (At2g46590), two closely related Arabidopsis genes (Figure 2A and Table 2). These two genes, display a high degree of sequence similarity and show identical patterns of expression, indicating a potential case of functional redundancy. Nevertheless, a systematic analysis of mutant variants demonstrated that they perform opposite functions in the control of seed germination [23–25]. Thus, *DAG1* and *DAG2* are clearly non-redundant and paralogous genes produced after a recent duplication event.

Conversely, phylogenetic relationship could help in the identification of gene function. Considering the case described above, a third gene (At4g24060) seems to be a paralog of the *DAG1-DAG2* branch, resulting in a cluster that appears to be ortholog to the rice gene cluster (OsDof-9/OsDof-18) present in MCOG *Cc*. Remarkably, OsDof-9 and OsDof-18 were first reported after their isolation from rice seed aleurone layers [27] and it will be interesting to investigate whether they have evolved into antagonistic functions in germination as their Arabidopsis corresponding paralogs.

Considering genome size differences in Arabidopsis (115 Mb) and rice (420 Mb), it is worth mentioning that 36 Dof genes were identified in the former, whereas only 30 in the later, in agreement with important duplication events in the origin of the Arabidopsis genome. In this context, establishing phylogenetic relationships is of outstanding interest to unravel gene functionality.

Conclusions

We identified the probable full complement of Dof genes in rice and Arabidopsis, which are representative of the major evolutionary lineages in the angiosperms: the monocotyledons and the dicotyledons. Phylogenetic analyses resulted in the identification of four major clusters of orthologous genes that contain members belonging to both species, and that must have been represented in their common ancestor before the taxonomic splitting of the angiosperms. Recognition of species-specific subgroups within these clusters led to explore the existence of monocot and dicot unique genes. We performed exhaustive searches in plant databases that allowed the detection of likely orthologs to dicot-specific genes, while no clear orthologs to monocot-specific-genes could be identified. In view of important genome duplication events leading to gene redundancy in the history of plant diversification, a combination of available functional data with phylogenetically inferred relationships are essential to effectively establish conserved and diverged roles in present day genes of evolutionary unrelated plant species.

Methods

Our collection of non-redundant Arabidopsis Dof proteins was gathered from three different and interconnected sources: the Munich Information Center for Protein Sequences database (MIPS, MATDB; http://mips.gsf.de/proj/thal/db), the Institute for Genomic Research, (*At* TIGR db, http://www.tigr.org/tdb/e2k1/ath1/index.shtml) and the Regulatory Gene Initiative on Arabidopsis (REGIA) European project. Information regarding the gene structure was obtained from the *At* TIGR db. Redundancy analyses of the Arabidopsis genomic regions comprising Dof genes were carried out with the Redundancy Viewer at the MATDB.

The compilation of a non-redundant set of rice Dof proteins was obtained from the *Oryza sativa* ssp. *japonica* and *Oryza sativa* ssp. *indica* databases. Sequences for *japonica* were obtained from the International Rice Genome Sequencing Project, IRGSP, through the Rice TIGR db BLAST tool http://www.tigr.org/tdb/e2k1/osa1/index.shtml. Newly released sequences from chromosomes 1 and 4 [33,34] were obtained from the DNA Data Bank of Japan http://www.ddbj.nig.ac.jp/Welcome-e.html. Additional *japonica* sequences were obtained from the Syngenta project by browsing the Torrey Mesa Research Institute (TMRI) Rice Genome Database http://www.tmri.org/. Sequences for *indica* were obtained from the Whole Genome Shotgun Sequencing Project of the Beijing Genomics Institute by means of the *O. sativa* BLAST page at the NCBI http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/riceWGS.html. Gene structure of previously annotated Dof genes was obtained from the Rice TIGR db. Unannotated Dof genes were annotated using the Rice Genome Automated Annotation System (RiceGAAS; http://ricegaas.dna.affrc.go.jp) at the National Institute of Agrobiological Science. Additional information was supplied from the MATDB.

Alignments of protein sequences by the CLUSTALW [43,45,46] were performed at the DNA Data Bank of Japan page http://www.ddbj.nig.ac.jp/Welcome-e.html. Bootstrapping analysis with a PHYLIP format tree output was carried out after the neighbor-joining method and the trees were represented with the help of the TREEVIEW (v. 1.6.6) software [47]. Rice and Arabidopsis conserved motif analysis within the determined Dof groups was performed by means of the RiceGAAS, MEME ([48]; http://meme.sdsc.edu/meme/website/intro.html) and BLOCKS ([49]; http://blocks.fhcrc.org/blocks) programs.

Author's Contributions

DL carried out the annotation of the rice genes, the phylogenetic, bioinformatic and genomic analyses, drafted and edited the manuscript. PC contributed with the Dof gene family background knowledge and edited the manuscript. JVC conceived of the study and participated in its design and coordination. All authors read and approved the final manuscript.

Acknowledgments

We are grateful to all the researchers working on Dof transcription factors for supplying the experimental data used in this study. We thank B. Burr & F. Burr (Brookhaven National Laboratory, USA) and J. Dopazo (Bioinformatics Unit, CNIO Spain) for helpful discussions and critical reading of this manuscript. This work was supported by grants from Comunidad Autonoma de Madrid (07B/0011/2002), Ministerio de Ciencia y Tecnología (BMC2000-1483) and EU-REGIA (QLRT-1999-00876), and by a postdoctoral fellowship from the Comunidad Autonoma de Madrid to D.L.

References

1. Tatusov RL, Koonin EV and Lipman DJ: **A genomic perspective on protein families** *Science* 1997, **278:**631-7.
2. Soullier S, Jay P, Poulat F, Vanacker JM, Berta P and Laudet V: **Diversification pattern of the HMG and SOX family members during evolution** *J Mol Evol* 1999, **48:**517-27.
3. The_Arabidopsis_Genome_Initiative: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana** *Nature* 2000, **408:**796-815.
4. The_C._elegans_Sequencing_Consortium: **Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology** *Science* 1998, **282:**2012-2018.
5. Thornton JW and DeSalle R: **Gene family evolution and homology: genomics meets phylogenetics** *Annu Rev Genomics Hum Genet* 2000, **1:**41-73.
6. Goffeau A: **The yeast genome** *Pathol Biol* 1998, **46:**96-7.
7. Dyson N: **The regulation of E2F by pRB-family proteins** *Genes Dev* 1998, **12:**2245-2262.
8. Boggon TJ, Shan WS, Santagata S, Myers SC and Shapiro L: **Implication of tubby proteins as transcription factors by structure-based functional analysis** *Science* 1999, **286:**2119-25.
9. Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T and Parcy F: **bZIP transcription factors in Arabidopsis** *Trends Plant Sci* 2002, **7:**106-11.
10. Singh KB: **Transcriptional regulation in plants: the importance of combinatorial control** *Plant Physiol* 1998, **118:**1111-20.
11. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ and Samaha RR *et al.*: **Arabidopsis tran-**

scription factors: genome-wide comparative analysis among eukaryotes *Science* 2000, **290:**2105-10.
12. Sasaki T and Burr B: **International Rice Genome Sequencing Project: the effort to completely sequence the rice genome** *Curr Opin Plant Biol* 2000, **3:**138-41.
13. Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y and Zhang X *et al.*: **A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica)** *Science* 2002, **296:**79-92.
14. Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P and Varma H *et al.*: **A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica)** *Science* 2002, **296:**92-100.
15. Barry GF: **The use of the Monsanto draft rice genome sequence in research** *Plant Physiol* 2001, **125:**1164-5.
16. Yanagisawa S: **A novel DNA-binding domain that may form a single zinc finger motif** *Nucleic Acids Res* 1995, **23:**3403-10.
17. Yanagisawa S: **A novel multigene family that the gene for a maize DNA-binding protein, MNB1a belongs to: isolation of genomic clones from this family and some aspects of its molecular evolution** *Biochem Mol Biol Int* 1996, **38:**665-73.
18. Yanagisawa S and Schmidt RJ: **Diversity and similarity among recognition sequences of Dof transcription factors** *Plant J* 1999, **17:**209-14.
19. Vicente_Carbajosa J, Moose SP, Parsons RL and Schmidt RJ: **A maize zinc-finger protein binds the prolamin box in zein gene promoters and interacts with the basic leucine zipper transcriptional activator Opaque2** *Proc Natl Acad Sci U S A* 1997, **94:**7685-90.
20. Mena M, Vicente_Carbajosa J, Schmidt RJ and Carbonero P: **An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm** *Plant J* 1998, **16:**53-62.
21. Yanagisawa S and Sheen J: **Involvement of maize Dof zinc finger proteins in tissue-specific and light-regulated gene expression** *Plant Cell* 1998, **10:**75-89.
22. Chen W, Chao G and Singh KB: **The promoter of a H2O2-inducible, Arabidopsis glutathione S-transferase gene contains closely linked OBF- and OBP1-binding sites** *Plant J* 1996, **10:**955-66.
23. Papi M, Sabatini S, Bouchez D, Camilleri C, Costantino P and Vittorioso P: **Identification and disruption of an Arabidopsis zinc finger gene controlling seed germination** *Genes Dev* 2000, **14:**28-33.
24. Papi M, Sabatini S, Altamura MM, Hennig L, Schafer E, Costantino P and Vittorioso P: **Inactivation of the phloem-specific Dof zinc finger gene DAG1 affects response to light and integrity of the testa of Arabidopsis seeds** *Plant Physiol* 2002, **128:**411-7.
25. Gualberti G, Papi M, Bellucci L, Ricci I, Bouchez D, Camilleri C, Costantino P and Vittorioso P: **Mutations in the Dof Zinc Finger Genes DAG2 and DAG1 Influence with Opposite Effects the Germination of Arabidopsis Seeds** *Plant Cell* 2002, **14:**1253-1263.
26. Mena M, Cejudo FJ, Isabel-Lamoneda I and Carbonero P: **A Role for the DOF Transcription Factor BPBF in the Regulation of Gibberellin-Responsive Genes in Barley Aleurone** *Plant Physiol* 2002, **130:**111-119.
27. Washio K: **Identification of Dof proteins with implication in the gibberellin-regulated expression of a peptidase gene following the germination of rice grains** *Biochim Biophys Acta* 2001, **1520:**54-62.
28. Kisu Y, Ono T, Shimofurutani N, Suzuki M and Esaka M: **Characterization and expression of a new class of zinc finger protein that binds to silencer region of ascorbate oxidase gene** *Plant Cell Physiol* 1998, **39:**1054-64.
29. Kisu Y, Harada Y, Goto M and Esaka M: **Cloning of the pumpkin ascorbate oxidase gene and analysis of a cis-acting region involved in induction by auxin** *Plant Cell Physiol* 1997, **38:**631-7.
30. Baumann K, De Paolis A, Costantino P and Gualberti G: **The DNA binding site of the Dof protein NtBBF1 is essential for tissue-specific and auxin-regulated expression of the rolB oncogene in plants** *Plant Cell* 1999, **11:**323-34.
31. Plesch G, Ehrhardt T and Mueller-Roeber B: **Involvement of TAAAG elements suggests a role for Dof transcription factors in guard cell-specific gene expression** *Plant J* 2001, **28:**455-64.
32. Vincentz M, Bandeira-Kobarg C, Gauer L, Schlogl P and Leite A: **Evolutionary Pattern of Angiosperm bZIP Factors Homologous to the Maize Opaque2 Regulatory Protein** *J Mol Evol* 2003, **56:**105-16.
33. Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y and Hu X *et al.*: **Sequence and analysis of rice chromosome 4** *Nature* 2002, **420:**316-20.
34. Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z and Nagamura Y *et al.*: **The genome sequence and structure of rice chromosome 1** *Nature* 2002, **420:**312-6.
35. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs** *Nucleic Acids Res* 1997, **25:**3389-402.
36. Yanagisawa S: **The Dof family of plant transcription factors** *Trends Plant Sci* 2002, **7:**555-60.
37. Graur D and Li W-H: **Fundamentals of molecular evolution** *Sunderland, MA: Sinauer Associates* 22000.
38. Hall AE, Fiebig A and Preuss D: **Beyond the Arabidopsis genome: opportunities for comparative genomics** *Plant Physiol* 2002, **129:**1439-47.
39. Simillion C, Vandepoele K, Van Montagu MC, Zabeau M and Van De Peer Y: **The hidden duplication past of Arabidopsis thaliana** *Proc Natl Acad Sci U S A* 2002, **99:**13627-32.
40. Vision TJ, Brown DG and Tanksley SD: **The origins of genomic duplications in Arabidopsis** *Science* 2000, **290:**2114-7.
41. Tatusov RL, Galperin MY, Natale DA and Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution** *Nucleic Acids Res* 2000, **28:**33-6.
42. Koonin EV, Tatusov RL and Galperin MY: **Beyond complete genomes: from sequence to structure and function** *Curr Opin Struct Biol* 1998, **8:**355-63.
43. Fitch WM: **Distinguishing homologous from analogous proteins** *Syst Zool* 1970, **19:**99-113.
44. Page RD and Charleston MA: **From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem** *Mol Phylogenet Evol* 1997, **7:**231-40.
45. Atchley WR and Fitch WM: **A natural classification of the basic helix-loop-helix class of transcription factors** *Proc Natl Acad Sci U S A* 1997, **94:**5172-6.
46. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W and Dress AW: **Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis** *Mol Biol Evol* 2000, **17:**164-78.
47. Page RD: **TreeView: an application to display phylogenetic trees on personal computers** *Comput Appl Biosci* 1996, **12:**357-8.
48. Bailey TL and Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers** *Proc Int Conf Intell Syst Mol Biol* 1994, **2:**28-36.
49. Henikoff S, Henikoff JG and Pietrokovski S: **Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations** *Bioinformatics* 1999, **15:**471-9.
50. Park DH, Lim PO, Kim JS, Cho DS, Hong SH and Nam HG: **The Arabidopsis COG1 gene encodes a Dof domain transcription factor and negatively regulates phytochrome signaling** *Plant J* 2003, **34:**161-71.